# What Goes Around Comes Around: Learning Sentiments in Online Medical Forums.

**Victoria Bobicev**
Technical University of Moldova
vika@rol.md

**Marina Sokolova**
University of Ottawa, Institute for Big Data Analytics, Canada
sokolova@uottawa.ca

**Michael Oakes**
Research Group in Computational Linguistics, University of Wolverhampton, UK
Michael.Oakes@wlv.ac.uk

## Abstract

Currently 19%-28% of Internet users participate in online health discussions. A 2011 survey of the US population estimated that 59% of all adults have looked online for information about health topics such as a specific disease or treatment. Although empirical evidence strongly supports the importance of emotions in health-related messages, there are few studies of the relationship between a subjective language and online discussions of personal health. In this work, we study sentiments expressed on online medical forums. As well as considering the predominant sentiments expressed in individual posts, we analyze sequences of sentiments in online discussions. Individual posts are classified into one of five categories. We identified three categories as sentimental (encouragement, gratitude, confusion) and two categories as neutral (facts, endorsement). 1438 messages from 130 threads were annotated manually by two annotators with a strong inter-annotator agreement (Fleiss kappa = 0.737 and 0.763 for posts in sequence and separate posts respectively). The annotated posts were used to analyse sentiments in consecutive posts. In four multi-class classification problems, we assessed HealthAffect, a domain-specific affective lexicon, as well general sentiment lexicons in their ability to represent messages in sentiment recognition.

## 1 Motivation

User-friendly Web 2.0 technologies encourage the general public to actively participate in the creation of the Web content. Blogs, social networks, and message boards reach out to a global community of Web users. These online texts discuss personal experience and convey sentiments and emotions of the authors. These emotion-rich posts are known to be important in setting interaction patterns among members of online communities, as emotion-rich text has a strong influence on a public mood (Allan, 2005). Studies of online sentiments and opinions can help in the understanding of sentiments and opinions of the public at large. Such understanding is especially important for the development of public policies whose success greatly depends on public support, e.g. education, health care, housing and infrastructure.

Effective implementation of health care policies relies on the understanding of opinions expressed by the general public. Major health care initiatives such as vaccination during pandemics and the incorporation of healthy choices in everyday life styles are examples of policies that require such understanding to be successfully implemented. As online media becomes the main medium for the posting and exchange of information, analysis of this online data can contribute to studies of the general public's opinions on health-related matters. Currently 19%-28% of Internet users participate in online health discussions. Surveys of medical forum participants revealed that personal testimonials attract attention of up to 49% of participants, whereas only 25% of participants are motivated by scientific and practical content (Balicco and Paganelli, 2011). Analysis of the information posted online contributes to the effectiveness of decisions on public health (Paul and Drezde, 2011; Chee et al., 2009). A 2011 survey of the US population estimated that 59% of all adults have looked online for information about health topics such as a specific disease or treatment (Fox 2011). Although empirical evidence strongly supports the importance of emotions in health-related messages (Pennebaker and Chung, 2006), there are few studies of the relationship between a subjective language and online discussions of personal health (Smith 2011).

Our interest concentrates on sequences of sentiments in the medical forum discourse. It has been shown that sentiments expressed by a forum participant affect sentiments in messages written by other participants posted on

the same discussion thread (Zafarani et al., 2010). Shared online emotions can improve personal well-being and empower patients in their battle against an illness (Malik and Coulson, 2010). We aimed to identify the most common sentiment pairs and triads and to observe their interactions. We applied our analysis to data gathered from the In Vitro Fertilization (IVF) medical forum.[1] This forum is designed to bring together women who use IVF treatments with the hope of conceiving. As a result, women constitute 95% of the forum participants and they post almost 99% of the messages, although there are occasional messages posted by men. To give a glimpse of the emotionally-charged data, we provide an example of four consecutive messages from an embryo transfer discussion:

*Alice*: Jane - whats going on??

*Jane:* We have our appt. Wednesday!! EEE!!!

*Beth*: Good luck on your transfer! Grow embies grow!!!!

*Jane*: The transfer went well - my RE did it himself which was comforting. 2 embies (grade 1 but slow in development) so I am not holding my breath for a positive. This really was my worst cycle yet; it was the Antagonist protocol which is supposed to be great when you are over 40 but not so much for me!!

In our sentiment analysis, we applied a three-fold approach. First, we manually annotated the messages, analyzed the dominant sentiments which appeared in the medical forum, and computed the agreement between the annotators. Second, we built a domain-specific lexicon HealthAffect for future sentiment analysis of the data. Third, we used HealthAffect in the automated sentiment recognition. We obtained the following results: we identified the dominant sentiments as *encouragement*, *gratitude*, *confusion*, *facts*, and *endorsement*. 1438 messages from 130 threads were annotated manually by two annotators with a strong inter-annotator agreement (Fleiss kappa = 0.737 and 0.763 for posts in sequence and separate posts respectively). The annotated posts were used to analyze sentiments in consecutive posts. In automated sentiment classification, we applied HealthAffect, a domain-specific lexicon of affective words. The current work is builds on results which appeared in (Bobicev et al, 2014).

The article is organized as follows: Section 2 presents relevant work in sentiment analysis, Section 3 introduces the data set, Section 4 - the annotation scheme and its results, Section 5 presents the correspondence and sequence results, Section 6 describes sentiment classification experiments, and Section 7 discusses the results.

## 2 Relevant Work

**Sentiment analysis.** The availability of emotion-rich text has helped to promote studies of sentiments from a boutique science into the mainstream of Text Data Mining (TDM). The "sentiment analysis" query on Google Scholar returns about 16,800 hits in scholarly publications appearing since 2010. In sentiment analysis, Machine Learning (ML) methods, affective lexicons, and Natural Language Processing (NLP) tools are used to classify text units (e.g., words, sentences, paragraphs) into sentiment categories (Taboada et al, 2011). A message is the core text unit in online forums. Hence we decided to keep it as our text unit.

Reliable annotation is essential for a thorough analysis of text. Multiple annotations of topic-specific opinions in blogs were evaluated by Osman et al. (2010). The authors computed agreement among seven manual annotators for five classification categories, including positive, negative, mixed opinions and non-opinionated and non-relevant categories. Sokolova and Bobicev (2013) evaluated annotation agreement achieved on messages gathered from a medical forum. Bobicev et al. (2012) used multiple annotators to categorize tweets into those positive, negative and neutral sentiments. The merits of reader-centric and author-centric annotation models were discussed in (Balahur, Steinberger, 2009). In this work, we apply the reader-centric annotation model. We use Fleiss Kappa (Nichols et al, 2010) to evaluate inter-annotator agreement.

An accurate sentiment classification relies on electronic sources of semantic information. Sentiment research often uses lexicons where words are assigned into opinion, sentiment, and emotion categories. However, in independent studies (Sokolova and Bobicev, 2013) and (Goeuriot et al, 2011), the authors showed that the sentiment categories of SentiWordNet[2], WordNetAffect[3] and the Subjectivity lexicon[4] are not fully representative of health-related emotions. We use HealthAffect, a domain-specific lexicon, to automatically classify sentiments. A preliminary, much smaller version of the lexicon was introduced in (Sokolova and Bobicev, 2013). In the current work, we re-populate the lexicon and use a manual filtering to prevent over-fitting the data.

[1] http://ivf.ca/forums
[2] http://sentiwordnet.isti.cnr.it/
[3] http://wndomains.fbk.eu/wnaffect.html
[4] http://mpqa.cs.pitt.edu/#subj_lexicon

Sentiment propagation is an emerging area in sentiment analysis. Although the relationship between consecutive sentiments is a popular subject of a fine-grained discourse analysis (Smith and Lee, 2014), it only recently started to make inroads into text mining. Subjective information posted by a user may affect subjectivity in posts written by other users (Zafarani et al 2010). Tsai et al (2013) used a two-step approach to evaluate sentiment propagation among related common sense concepts. Correlations between emotions expressed in consecutive posts were studied in (Chmiel et al, 2011; Tan et al, 2011; Hassan et al, 2012). On the other hand, health-related sentiment classification has focused on individual messages. Our current work goes beyond individual messages and studies sequences of sentiments in consecutive posts.

**Concept-level sentiments.** Our approach is reminiscent of concept-level sentiment analysis (Cambria, 2013). In the analysis of data, we retrieve and aggregate subjective information about different aspects of IVF treatment. Such information is directly linked with the basic IVF concepts and features, thus, cannot be identified through a keyword search or the use of general lexical resources.

Another technique associated with concept-level analysis is correspondence analysis, a multivariate technique for analyzing matrices of data. Its implementation in the R programming language is described by Baayen (2008). The technique of correspondence analysis discovers whether groups of words tend to occur in the same messages as each other. Such groups are called "factors", and they are ordered according to their importance in terms of how much of the variation between the messages they explain. The idea for such a representation comes from work by Stanley and Meyer (2009), who used another matrix analysis technique called Factor Analysis to plot students' ratings of their emotional states on various occasions on a two-dimensional graph. Stanley and Meyer call the discovered axes (and hence constructs) for representing affective experiences "affective space". We applied correspondence analysis in our study of sentiments.

**Reproductive technologies and sentiments.** Reproductive technologies belong to a group of hotly debated health care issues in modern society. These highly spirited debates are in part due to a multitude of issues connected with the technologies. For example, the most popular reproductive technology - In Vitro Fertilization - is linked to an uncertain chance of live birth and discussions of the health of the babies born, ongoing pregnancies, clinical pregnancies, miscarriages, multiple pregnancies, implantation rate, cryopreservation rate, embryo quality and fertilization rate (Mantikou et al, 2013), as well as age, obesity, a risk of breast cancer and overall financial costs to society (Pantasri and Norman, 2013). The complexity of the problem causes the technology's recipients to seek information, advice, and guidance not only from medical professionals, but from peers as well. The peer connection is increasingly done online, through social media (Zillen, 2011).

A meta-study of 19 studies on reproductive technologies published in 1999-2009 listed several reasons for the use of medical forums: a) information searching - to learn about psychological, physical and social aspects of available treatments, evaluations of alternative treatments; b) in seeking emotional support - anonymous communication, immediate and constant community access, easy contact to peers (Zillen, 2011). In a manual survey of online infertility support groups, empathy and shared personal experience constituted 45.5% of content, gratitude - 12.5%, recognized friendship with other members - 9.9%, whereas the provision of information and advice and requests for information or advice took up 15.9% and 6.8% respectively (Malik and Coulson, 2010). Our analysis supports that observation with the results obtained on a large number of messages.

Sentiment analysis often connects its subjects with specific online media (e.g., sentiments on consumer goods are studied on Amazon.com). Health-related emotions are studied on Twitter (Chew and Eysenbach, 2010; Bobicev et al, 2012) and online public forums (Malik and Coulson, 2010; Goeuriot et al, 2012). In this work, we continue studies of online forum data.

## 3 The IVF Forum Data

We worked with online messages posted on a medical forum. The forum communication model promotes messages which disclose the emotional state of the authors. We gathered data from the In Vitro Fertilization (IVF) website dedicated to reproductive technologies, a hotly debated issue in modern society. The website belongs to an infertility outreach resource community created by prospective, existing and past IVF patients. The IVF.ca website includes forums: *Cycle Friends*, *Expert Panel*, *Trying to Conceive*, *Socialize*, *In Our Hearts*, *Pregnancy*, *Parenting*, and *Administration*.[5] Every forum hosts a few sub-forums, e.g. the *Cycle Friends* forum has six sub

[5] www.ivf.ca/forums

forums: *Introductions*, *IVF/FET/IUI Cycle Buddies*, *IVF Ages 35+*, *Waiting Lounge*, *Donor & Surrogacy Buddies*, and *Adoption Buddies*. On every sub-forum, new topics are initiated by the forum participants. Depending on the interest among participants, a different number of messages is associated with each topic, e.g., *Human growth hormone & what to expect* has 120 messages posted from Oct 2012, while *Over 40 and pregnant or trying to be* has 3,455 messages posted from May 2010.

We wanted the forum to represent a variety of discussions and contain a manageable number of topics and messages. The *IVF Ages 35+* sub-forum[6] satisfied both requirements, i.e., it had 510 topics and 16388 messages, where the messages had 128 words on average[7]. Figure 1 illustrates the distribution of posts among the forum topics.

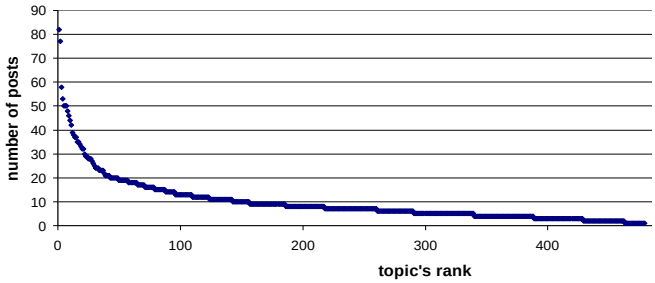


Figure 1: Number of posts per topic in the *IVF Ages 35+* sub-forum

All topics were initiated by the forum participants. Among those, 340 topics contained < 10 posts. These short topics often contained one initial request and a couple of replies and were deemed too short to form a good discussion. We also excluded topics containing > 20 posts. This exclusion left 80 topics with an average of 17 messages per topic for a manual analysis by two annotators.
The topics usually had the following structure:

a) a participant started the theme with a post;
   i) the initial post usually contained some information about the participant's problem, expressed worry, concern, uncertainty and a request for help to the other forum participants;
b) the following posts:
   i) provided the requested information by describing their similar stories, knowledge about treatment procedures, drugs, doctors and clinics, or
   ii) supplied moral support through compassion, encouragement, wishing all the best, good luck, etc.
c) the participant who started the topic often thanked other contributors and expressed appreciation for their help and support.

We have focused on recognition of sentiment sequences expressed in the IVF forum data. We wanted to identify types of sentiments dominant in these messages and how these sentiments influence each other.

[6] http://ivf.ca/forums/forum/166-ivf-ages-35/
[7] We harvested the data in July 2012.

## 4 Data Annotation

Annotation of subjectivity can be centered either on the perception of a reader (Strapparava and Mihalcea, 2008) or the author of a text (Balahur and Steinberger, 2009). In the current work, we aimed to detect sentiments conveyed by posts on the discussion readers. Hence, we opted for the reader perception model and asked annotators to analyze the topic's sentiment as it was addressed to the other forum participants. The data annotation was carried out by the Master's students as their practical work for the course "Semantic Interpretation of Text". The students had already completed courses on "Computational Linguistics" and "Natural Language Processing". Based on the quality of annotations, eight annotators were selected after the first phase of the sentiment analysis. Most annotators already had experience in sentiment and opinion annotation. Each annotator independently annotated a set of topics. Each message was annotated by two annotators.

We used 292 randomly selected posts to verify whether the messages were self-evident for sentiment annotation or required an additional context. The annotators reported that posts were long enough to convey emotions and in most cases there was no need for a wider context.

We applied an annotation scheme which was successfully applied in (Sokolova and Bobicev, 2013). In (Malik, Coulson, 2010), the authors showed that most posts referred to sharing personal experiences, provision of information or advice, expressions of gratitude/friendship, chat, requests for information, and expressions of universality (e.g. "we're all in this together"). Hypothesizing that binary sentiment categories (e.g., positive and negative polarity) would be too general and could not adequately cover emotions expressed in health-related messages, we intended to build a set of sentiments that

1. contains sentiment categories specific for posts from medical forums, and
2. makes an automated sentiment detection feasible and reliable.

We used the bottom up approach to build that set. First, we asked annotators to read several topic discussions and describe sentiments expressed by the forum participants and the sentiment propagation within these discussions. More specifically, the annotators were told to indicate sentiments in sequences. For example, we asked annotators to answer groups of questions:

- What sentiment was expressed in the first post in the topic? How were the sentiments of the following posts affected by the initial sentiment?
- How long did an expressed sentiment last in the topic? If it was replaced by another one, how did the replacement happen?
- Did the participants joining the discussion try to change the previous sentiments? Did the participants succeed in such attempts?

We told annotators that they do not to mark descriptions of symptoms and diseases as subjective; in many cases they appear in the post as objective information for other forum participants that have encountered similar issues. In such cases only the author's sentiments toward other participant should be taken into consideration. For example, `I have had a few days now with heartburn/reflux - could be stress, a little achy tummy/pelvic and a tired aching back. More waiting, but getting more hopeful` is a description of symptoms and should not be annotated as subjective. In contrast, `I hope your visit with us infertilies is short and sweet and you get that baby soon!!!` exposes the author's sentiment towards another person.[8] It should be mentioned that the posts were usually long enough to express several sentiments. However, annotators were requested to mark messages with one sentiment category.

After gathering results of the initial annotation, we merged and summarized the annotations. That resulted in 35 sentiment types which we placed into three groups:

- ***confusion***, which included worry, concern, doubt, impatience, uncertainty, sadness, anger, embarrassment, hopelessness, dissatisfaction, and dislike;
- ***encouragement***, which included cheering, support, hope, happiness, enthusiasm, excitement, optimism;
- ***gratitude***, which included thankfulness.

A special group of sentiments was presented by expressions of compassion, sorrow, and pity. According to the WordNetAffect classification, these sentiments should be considered negative. However, in the context of

[8] All examples preserve original spelling and grammar.

health discussions, these emotional expressions appeared in conjunction with moral support and encouragement. Hence, we treated them as a part of *encouragement.*

Not all posts had an emotional content. Posts presenting only factual information were marked as *facts*. Some posts contained factual information and strong emotional expressions; those expressions almost always conveyed encouragement ("*hope, this helps*", "*I wish you all the best*", "*good luck*"). Such posts were labeled *endorsement.* Note that the final categories did not manifest negative sentiments. In lieu of negative sentiments, we considered *confusion* as a non-positive label. *Encouragement* and *gratitude* were considered positive labels, *facts* and *endorsement* - neutral.

The posts that both annotators labelled with the same label were assigned to this category; 1256 posts were assigned with a class label. The posts labelled with two different sentiment labels were marked as *ambiguous*; 182 posts were marked as *ambiguous.*

We evaluated agreement between the annotators by using Fleiss Kappa (Nichols et al, 2010), a measure that evaluates agreement for a multi-class manual labeling.

*Fleiss Kappa* = $(P - P_{class})/(1-P_{class})$

where P is an average agreement per a post and $P_{class}$ is an average agreement per a class.

Despite the challenging data, we obtained Fleiss Kappa = 0.737 which indicated a strong agreement between annotators (Osman et al, 2010). This value was obtained on 80 annotated posts. Agreement for the randomly extracted posts was calculated separately in order to verify whether annotation of separate posts was no more difficult than annotation of the post sequences. Contrary to our expectations, the obtained Fleiss Kappa = 0.763 was slightly higher than on the posts in discussions. The final distribution of posts among sentiment classes is presented in Table 2.

| **Classification category** | **# of posts** | **Per-cent** |
|---|---|---|
| *Facts* | 494 | 34.4% |
| *Encouragement* | 333 | 23.2% |
| *Endorsement* | 166 | 11.5% |
| *Confusion* | 146 | 10.2% |
| *Gratitude* | 131 | 9.1% |
| *Ambiguous* | 168 | 11.7% |
| Total | 1438 | 100% |

**Table 2:** Class distribution of the IVF posts.

## 5 Correspondence Analysis for Sentiment Sequences

We applied correspondence analysis (Baayen, 2008) to recognize the affective groups of the most frequent words found in the data. We used the messages from the ART_over_35 topic, missing out only the very short ones. The messages are numbered in the order they appear in the discussion. As input, we produced a matrix where the columns corresponded to the 500 most frequent words in the ART_over_35 text collection, and the rows each corresponded to one individual message. Since we were mainly interested in sentiment words, this original matrix was reduced by retaining only those columns corresponding to the 41 words conveying sentiments such as "best", "better" and "congratulations". From the list, 28 words are indicative of sentiment categories and appear in HealthAffect (e.g, able, against, interested, recommended, risk) and 13 words are not indicative of specific categories, thus do not appear in HealthAffect (e.g., avoid, luxury).

The technique of correspondence analysis discovers whether groups of words tend to occur in the same messages as each other. Such groups are called "factors", and they are ordered according to their importance in terms of how much of the variation between the messages they explain. The graph below (Figure 2) was produced by correspondence analysis, and shows to what extent each word and each message is related to the two main factors. Only those words which are significantly associated with the factors ($p < 0.1$) are shown in the graph. The group of words making up the first factor explain 24.5% of the variation between the posts, while those making

up the second factor explain 12.3% of this variation. The identified words occur together in three main groups: *concern, support and good will*, and *desire to know*. The groups form the affective *author-centric* space of the topic and can be representative of the affective space of the IVF discussion (Stanley and Meyer, 2009).

The graph shows that in the top left quadrant are words which occur together in messages expressing *concern* for the future, as in "I don't feel able to handle the negative pressure". In the top right quadrant are words which appear in messages of *support and good will*, such as "successful", "luck" and "good". Finally, in the lower left quadrant are words found in messages expressing a *desire to know*, "interested", "confusion", "success" and "like" (as in "I'd like to know the chances of success"). Most of the early messages (from 1 to 17) are in the topic-opening "desire to know" quadrant, apart from a short exchange of *anxious* messages (10, 12 and 13). There are then a series of encouraging messages (from 18 to 27), while the last few messages are more neutral, scoring about 0 on Factor 2, and slightly negative on Factor 1. Although this is not apparent from the graph, they correspond to messages where people looked back on their own experiences of IVF in a neutral, unemotional way.

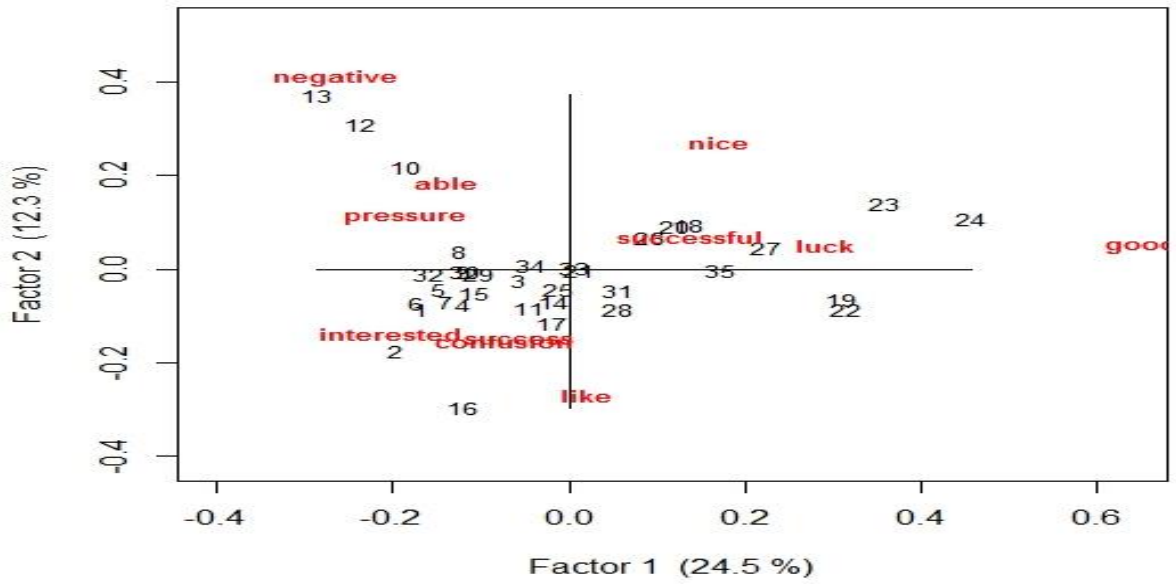


**Figure 2:** Correspondence analysis of sentiments

Note that there are no significant words in the fourth quadrant. We show only the words which were significantly associated with the factors ($p < 0.1$), and there are none of these in the fourth quadrant. Although messages 19 and 22 are in the fourth quadrant, the most important thing is that they score highly on Factor 1 (i.e. over to the right hand side of the graph). The last few messages (29 to 35) are not in the fourth quadrant, but appear about half way up (very close to the horizontal axis) mostly on the left.

To further identify sentiments that reinforced themselves and sentiments that were likely to trigger changes, we computed the distribution of sentiment pairs and triads in consecutive messages. We found that the most frequent sequences consisted mostly of *facts* and/or *encouragement*: 39.5% in total. These two categories were most likely to propagate through next messages. The most frequent change was from *endorsement* to *facts* (6.1% in total). Approximately 10% of sentiment pairs are *factual* and/or *encouragement* followed by *gratitude*. *Confusion* was followed by *facts* and *encouragement* in 80% of cases. The most frequent triad containing *confusion* was *confusion, facts, facts.* That sentiment transition shows a high level of support among the forum participants. Other less frequent sequences appear when a new participant added her post in the flow. Tables 3 and 4 list the results.

| Sentiment pairs | Occurrence | Percent |
|---|---|---|
| *facts, facts* | 170 | 19.5% |
| *encouragement, encouragement* | 119 | 13.7% |
| *facts, encouragement* | 55 | 6.3% |
| *endorsement, facts* | 53 | 6.1% |
| *encouragement, facts* | 44 | 5.1% |

**Table 3:** The most frequent sequences of two sentiments and their occurrence in the data.

| Sentiment triads | Occurrence | Percent |
|---|---|---|
| *factual, factual, factual* | 94 | 12.8% |
| *encouragement, encouragement, encouragement* | 63 | 8.6% |
| *encouragement, gratitude, encouragement* | 18 | 2.4% |
| *factual, endorsement, factual* | 18 | 2.4% |
| *confusion, factual, factual* | 17 | 2.3% |

**Table 4:** The most frequent triads of sentiments and their occurrences in the data.

Our next goal was to build a tool for reliable identification of the sentiments in a large number of texts. This tool has to be general enough to allow for diversity of natural language expressions appearing in the forum data and exhaustive enough to recognize opinions expressed towards the IVF treatment.

## 6 Automated Sentiment Recognition

The first stage of our study identified that the forum messages belonged to 5 sentimental and neutral categories. For automated sentiment classification, we tested a multi-categorical WordNet-Affect (Strapparava et al, 2006), and SentiWordNet (Baccianella et al. 2010), Bing Liu's Opinion Lexicon (Liu, 2010), and the MPQA subjectivity lexicon (Wiebe et al., 2005) which recognize only positive and negative polarity of their terms. We also tested several lexicons with sentiment information which were announced recently: The freely-downloadable SentiStrength sentiment analysis program (Thelwall et al., 2012) contains the list of English words which express emotions, SenticNet 3 (Cambria, Hussain, 2012) is a knowledge base which contains information about the semantics and sentics associated with multi-word expressions, and DepecheMood (Staiano, Guerini, 2014) which contains more than 37 500 terms which were assigned numerical values representing degrees of 8 sentiment categories: afraid, amused, angry, annoyed, dont_care, happy, inspired, sad. This very large lexicon was *crowdsourced* from rappler.com news articles along with the information displayed by Rappler's Mood Meter, a small interface offering the readers the opportunity to click on the emotion that a given news article made them feel. This way, numerous votes have been collected and document-by-emotion matrix was built which was transformed into a word – emotion matrix, e.g., concerned - 0.129322883 AFRAID, 0.100615215 AMUSED, 0.170474974 ANGRY, 0.161903853 ANNOYED, 0.120271172 DONT_CARE, 0.108064155 HAPPY, 0.098734566 INSPIRED, and 0.110613182 SAD. Among the listed lexicons, the following lexicons were found fairly often in our data: SentiWordNet - 3 725 terms, MPQA - 1 418, SenticNet 3 - 1 342, SentiStrength - 1 131, and DepecheMood - 4 467. We used those terms to represent our data in Machine Learning experiments.

We also used the domain-specific lexicon HealthAffect introduced in (Sokolova and Bobicev, 2013). To build the lexicon, we adapted the Pointwise Mutual Information (PMI) approach (Turney, 2002):

$$PMI(word1, word2) = \log_2(p(word1 \ \& \ word2)/( p(word1) \ p(word2)))$$

The initial candidates consisted of unigrams, bigrams and trigrams of words with frequency $\geq 5$ appearing in unambiguously annotated posts (i.e., we omitted posts marked as uncertain). This was a list of candidates to be included in our HealthAffect lexicon. Note that the Part-of-Speech tagging used by Turney would be ineffective due to a high volume of textual noise (e.g., incomplete sentences, InternetSpeak, informal grammar). Next, for each class and each candidate, we calculated PMI(*candidate, class*) as

$$PMI(candidate, class) = \log_2( p(candidate \text{ in } class)/( p(candidate) \ p(class))).$$

Next, we calculated Semantic Orientation (SO) for each candidate and for each class as

$$SO\ (candidate, class) = PMI(candidate, class) - \Sigma\ PMI(candidate, other\_classes)$$

where *other_classes* include all the classes except the class that Semantic Orientation is calculated for. After all the possible SO were computed, each HealthAffect candidate was assigned with the class that corresponded to its maximum SO. Consequently, each candidate was considered an indicator of the class that provided it with the maximum SO. It should be noted that each class got different numbers of indicative candidates.

Domain-specific lexicons can be prone to data over-fitting (since, for example, they might contain personal and brand names). To avoid the over-fitting pitfall, we manually reviewed and filtered out non-relevant elements, such as personal and brand names, geolocations, dates, stop-words and their combinations (since_then, that_was_the, to_do_it, so_you). Table 5 presents the HealthAffect profile. Note that we do not report the *endorsement* profile as it combines *facts* and *encouragement.*

| Class | unigrams | Bigrams | trigrams | total | Examples |
|---|---|---|---|---|---|
| *Facts* | 204 | 254 | 78 | 536 | round_of_ivf, hearbeat, a_protocol |
| *Encouragement* | 127 | 107 | 68 | 302 | congratula-tions, is_hard, on-ly_have_one |
| *Confusion* | 63 | 143 | 34 | 240 | crying, away_from, any_of_you |
| *Gratitude* | 37 | 51 | 34 | 122 | appreciate, a_huge, thanks_for_your |

**Table 5:** Statistics of the HealthAffect lexicon.

In the Machine Learning experiments, we represented the messages by the lexicon terms. The classification's performance was evaluated through four multiclass classification results:

- 6-class classification where all 1 438 posts are classified into 6 classes, including ambiguous.
- 5-class classification where 1269 unambiguous posts are classified into the 5 classes.
- 4-class classification where all 1269 unambiguous posts are classified into *encouragement, gratitude, confusion,* and neutral (i.e., *facts* and *endorsement*), and
- 3-class classification into positive (*encouragement, gratitude*), negative (*confusion),* and neutral (*facts, endorsement*).

We applied Naive Bayes (NB), NB Text, NB multinomial, SVM, Decision Trees and KNN. Decision Trees and KNN performed considerably worse than other algorithms and we do not report their results. To select the best classifier, we used 10-fold cross-validation and computed *F-score (F).* We used the majority class baseline. The classification results are shown in Tables 6 – 9.

**Table 6:** Classification results for 6 classes, the baseline = 0.171. The best F-score is **0.491**, the 2nd best – ***0.432.***

| Lexicon | features | NB | DMNBtext | NBMultinomial | SVM |
|---|---|---|---|---|---|
| SentiWordNet | 3 725 | 0.322 | 0.424 | 0.385 | 0.415 |
| MPQA | 1 418 | 0.313 | 0.388 | 0.394 | 0.389 |
| SenticNet 3 | 1 342 | 0.326 | 0.399 | 0.393 | 0.402 |
| SentiStrenght | 1 131 | 0.335 | 0.407 | 0.394 | 0.414 |
| DepecheMood | 4 467 | 0.320 | ***0.432*** | 0.384 | 0.425 |
| HealthAffect | 1 189 | 0.402 | 0.484 | **0.491** | 0.432 |

**Table 7**: Classification results for 5 classes, the baseline = 0.215. The best F-score is **0.582**, the 2nd best – ***0.519.***

| Lexicon | features | NB | DMNBtext | NBMultinomial | SVM |
|---|---|---|---|---|---|
| SentiWordNet | 3 725 | 0.388 | 0.518 | 0.453 | 0.505 |
| MPQA | 1 418 | 0.385 | 0.476 | 0.463 | 0.475 |
| SenticNet 3 | 1 342 | 0.379 | 0.461 | 0.459 | 0.471 |
| SentiStrenght | 1 131 | 0.401 | 0.484 | 0.480 | 0.485 |
| DepecheMood | 4 467 | 0.366 | ***0.519*** | 0.454 | 0.507 |
| HealthAffect | 1 189 | 0.481 | 0.580 | **0.582** | 0.530 |

**Table 8:** Classification results for 4 classes, the baseline = 0.353. The best F-score is **0.667**, the 2nd best – ***0.618.***

| Lexicon | features | NB | DMNBtext | NBMultinomial | SVM |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| SentiWordNet | 3 725 | 0.507 | 0.611 | 0.552 | 0.594 |
| MPQA | 1 418 | 0.472 | 0.55 | 0.556 | 0.577 |
| SenticNet 3 | 1 342 | 0.496 | 0.557 | 0.556 | 0.571 |
| SentiStrenght | 1 131 | 0.498 | 0.566 | 0.553 | 0.578 |
| DepecheMood | 4 467 | 0.511 | ***0.618*** | 0.552 | 0.606 |
| HealthAffect | 1 189 | 0.597 | 0.657 | **0.667** | 0.607 |

**Table 9:** Classification results for 3 classes, the baseline = 0.353. The best F-score is **0.697**, the 2nd best – ***0.675.***

| Lexicon | features | NB | DMNBtext | NBMultinomial | SVM |
|---|---|---|---|---|---|
| SentiWordNet | 3 725 | 0.584 | 0.665 | 0.651 | 0.64 |
| MPQA | 1 418 | 0.553 | 0.645 | 0.618 | 0.623 |
| SenticNet 3 | 1 342 | 0.574 | 0.643 | 0.631 | 0.629 |
| SentiStrenght | 1 131 | 0.571 | 0.63 | 0.617 | 0.622 |
| DepecheMood | 4 467 | 0.588 | ***0.675*** | 0.663 | 0.660 |
| HealthAffect | 1 189 | 0622 | 0.672 | **0.697** | 0.656 |

The classification improved as we reduced the number of classes, hence reduced the uncertainty for the algorithms. At the same time, there was a remarkable correspondence in the performance, namely the 2nd best results were always provided by DepecheMood. We hypothesize that the ability of DepecheMood to recognize several sentiments instead of just positive and negative was critical. Among the algorithms, NB always outperformed SVM, although various versions performed better on various lexicons: DBNBtext was always the best with DepecheMood, NBMultinomial always the best with HealthAffect. The results obtained on the HealthAffect features were the best in all experiments.

For each sentiment class, our results were as follows:

- The most accurate classification occurred for *gratitude*. It was correctly classified in 83.6% of its occurrences. It was most commonly misclassified as *encouragement* (9.7%). Posts classified as *gratitude* are mostly the shortest ones containing only some words of gratitude and appreciation of others' help. As they usually do not contain any more information than this, there were fewer chances for them to be misclassified.
- The second most accurate result was achieved for *encouragement.* It was correctly classified in 76.7% of cases. It was misclassified as neutral (9.8%) because the latter posts contained some encouraging with the purpose of cheering up the interlocutor.
- The overall biggest misclassification occurred into a fact class. There are two reasons for that: first, it is the biggest class in the data, second, even the post was marked as encouragement, confusion or gratitude still contained some factual information. There were just a few posts which expressed only sentiments without any description of which facts ledup to them.

## 7 Discussion and Future Work

We have presented results of sentiment recognition in messages posted on a medical forum. Sentiment analysis of online medical discussions differs considerably from polarity studies of consumer-written product reviews, financial blogs and political discussions. While in many cases positive and negative sentiment categories are powerful enough, such a dichotomy is not sufficient for medical forums. We formulate our medical sentiment analysis as a multi-class classification problem in which posts were classified into *encouragement, gratitude, confusion, facts* and *endorsement*. We have run four multi-class sentiment classification problems on which we compared performance of ML algorithms and ability of sentiment lexicons to represent the data. We have shown that Naïve Bayes provides reliable sentiment classification. DepecheMood and Health Affect were more successful in the data representation than other lexicons.

In spite of sentiment annotation being highly subjective, we obtained a strong inter-annotator agreement between two independent annotators (i.e., Fleiss Kappa = 0.73 for posts in discussions and Fleiss Kappa = 0.76 for separate posts). The Kappa values demonstrated an adequate selection of classes of sentiments and appropriate annotation guidelines. However, many posts contained more than one sentiment in most cases mixed with some factual information. The possible solutions in this case would be (a) to allow multiple annotations for each post; (b) to annotate every sentence of the posts. A specific set of sentiments on the IVF forum suggested that we ap-

plied the PMI approach to build a domain-specific lexicon HealthAffect and then manually reviewed and generalized it.

In the current work we analyzed message sequences in order to reveal patterns of sentiment interaction. Manual analysis of a sample of data showed that topics contained a coherent discourse. Some unexpected shifts in the discourse flow were introduced by a new participant joining the discussion. In future work, we may include the post's author information in the sentiment interaction analysis. The information is also important for analysis of influence, when one participant is answering directly to another one citing in many cases the post which she answered to. Identifying sentiment propagation among related semantic concepts is another venue of the future work.

We plan to use the results obtained in this study for analysis of discussions related to other highly debated health care policies. One future possibility is to construct a Markov model for the sentiment sequences. However, in any online discussion there are random shifts and alternations in discourse which complicate application of the Markov model.

In the future, we aim to annotate more text, enhance and refine HealthAffect, and use it to achieve reliable automated sentiment recognition across a spectrum of health-related issues.

## References


Allan, K. 2005. *Explorations in Classical Sociological Theory*: Seeing the Social World. Pine Forge Press, 2005

Baayen, H. 2008. *Analysing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press.

Balicco, L., C. Paganelli. 2011. *Access to health information: going from professional to public practices*, Information Systems and Economic Intelligence: 4th International Conference - SIIE'2011.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. *SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.* Proceedings of the 7th Conference on International Language Resources and Evaluation, 2200-2204.

Bobicev, V., M, Sokolova, Y. Jaffer, D. Schramm. 2012. *Learning Sentiments from Tweets with Personal Health Information.* Proceedings of Canadian AI 2012, p.p. 37–48, Springer.

Bobicev, V., M. Sokolova, M. Oakes. 2014. *Recognition of Sentiment Sequences in Online Discussions*, Social-NLP-COLING.

Cambria, E. and A. Hussain. 2012. *Sentic Computing: Techniques, Tools, and Applications.* Springer.

Cambria, E. 2013. *An Introduction to Concept-Level Sentiment Analysis*, Proceedings of MICAI 2013, pp. 478–483, Springer.

Cambria, E., D. Olsher, and D. Rajagopal. 2014. SenticNet 3: A Common and Common-Sense Knowledge Base for Cognition-Driven Sentiment Analysis. Proceedings of AAAI 2014,.

Chee, B., R. Berlin, B. Schatz. 2009. *Measuring Population Health Using Personal Health Messages*. Proceedings of AMIA Symposium, 92 - 96.

Chew, C. and G. Eysenbach. 2010. *Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak*. PLoS One, 5(11).

Chmiel, A., J. Sienkiewicz, M. Thelwall, G. Paltoglou, K. Buckley, A. Kappas, J. Holyst. 2011. *Collective Emotions Online and Their Influence on Community Life*. PLoS one.

Goeuriot, L., J. Na, W. Kyaing, C. Khoo,Y. Chang, Y. Theng and J. Kim. 2012. *Sentiment lexicons for health-related opinion mining*. Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, p.p. 219 – 225, ACM.

Fox, S. 2011. *The Social Life of Health Information*. Pew Research Center's Internet & American Life Project,http://pewinternet.org/Reports/2011/Social-Life-of-Health-Info.aspx

Hassan, A., A. Abu-Jbara, D. Radev. 2012. *Detecting subgroups in online discussions by modeling positive and negative relations among participants*. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 59-70).

Liu B. 2010. *Sentiment Analysis and Subjectivity*. Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010.

Malik S. and N. Coulson. 2010. *Coping with infertility online: an examination of self-help mechanisms in an online infertility support group.* Patient Educ Couns, vol. 81, no. 2, pp. 315–318

Mantikou, E., M.A.F.M. Youssef , M. van Wely, F. van der Veen, H.G. Al-Inany, S. Repping and S. Mastenbroek. 2013. *Embryo culture media and IVF/ICSI success rates: a systematic review*, Human Reproduction Update, vol. 19, issue 3, pp. 210-220.

Nichols, T., P. Wisner, G. Cripe, and L. Gulabchand. 2010. *Putting the Kappa Statistic to Use.* Qual Assur Journal, 13, p.p. 57-61.

Osman, D., J. Yearwood, P. Vamplew. 2010. *Automated opinion detection: Implications of the level of agreement between human raters.* Information Processing and Management, 46, 331-342.

Paul, M. and M. Dredze. 2011. *You Are What You Tweet: Analyzing Twitter for Public Health*. Proceedings of ICWSM.

Pantasri, T. and R. J. Norman. 2013. *The effects of being overweight and obese on female reproduction: a review*. Gynecological Endocrinology, vol. 30, issue 2, pp. 90-94.

Pennebaker, J. and Chung, C. 2006. *Expressive Writing, Emotional Upheavals, and Health*. Handbook of Health Psychology, Oxford University Press.

Smith, C. 2011. *Consumer language, patient language, and thesauri: A review of the literature*. Journal of the Medical Library Association, 99(2), 135– 144, 2011.

Smith, P. and M. Lee. 2014. *Acknowledging Discourse Function for Sentiment Analysis*. Proceedings of CI-CLing.

Sokolova, M. and V. Bobicev. 2013. *What Sentiments Can Be Found in Medical Forums*? Recent Advances in Natural Language Processing, 633-639

Staiano, J.,and Guerini, M. 2014. *DepecheMood: a Lexicon for Emotion Analysis from Crowd-Annotated News.* Proceedings of ACL-2014.

Stanley, D. J. and J. P. Meyer. 2009. *Two-dimensional affective space: a new approach to orienting the axes*. Emotion, vol. 9, issue 2, pp. 214-237.

Strapparava, C. and R. Mihalcea. 2008. *Semeval-2007 task 14: Affective text.* Proceedings of the 2008 ACM symposium on Applied computing, 2008.

Strapparava, C., A. Valitutti, and O. Stock. 2006. *The affective weight of the lexicon.* Proceedings of the 5th International Conference on Language Resources and Evaluation, pages 474-481, 2006.

Tan, C., L. Lee , J. Tang , L. Jiang , M. Zhou, P. Li, 2011. *User-level sentiment analysis incorporating social networks*, Proceedings of the 17th ACM SIGKDD international conference on KDDM.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll and M. Stede. 2011. *Lexicon-Based Methods for Sentiment Analysis*. Computational Linguistics 37 (2): 267-307.

Thelwall, M., Buckley, K.,and Paltoglou, G. 2012. *Sentiment strength detection for the social Web*. Journal of the American Society for Information Science and Technology, 63(1), 163-173.

Tsai, A., Tsai, R., Hsu, J., 2013. *Building a concept-level sentiment dictionary based on commonsense knowledge*. IEEE Intelligent Systems 28(2), 22–30.

Turney, P.D. 2002. *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*. Proceedings of ACL'02, Philadelphia, Pennsylvania, pp. 417-424.

Wiebe, Janyce; Theresa Wilson; and Claire Cardie. 2005. *Annotating expressions of opinions and emotions in language*. Language Resources and Evaluation 39: 165-210.

Zafarani, R., W. Cole, and H. Liu. 2010. *Sentiment Propagation in Social Networks: A Case Study in LiveJournal*. Advances in Social Computing (SBP 2010), pp. 413–420, Springer.

Zillen, N. 2011. Internet Use of Fertility Patients: A Systemic Review of the Literature. *Journal of Reproductive Medicine and Endocrinology*, 8(4): 281–287